Python Machine learning Labs

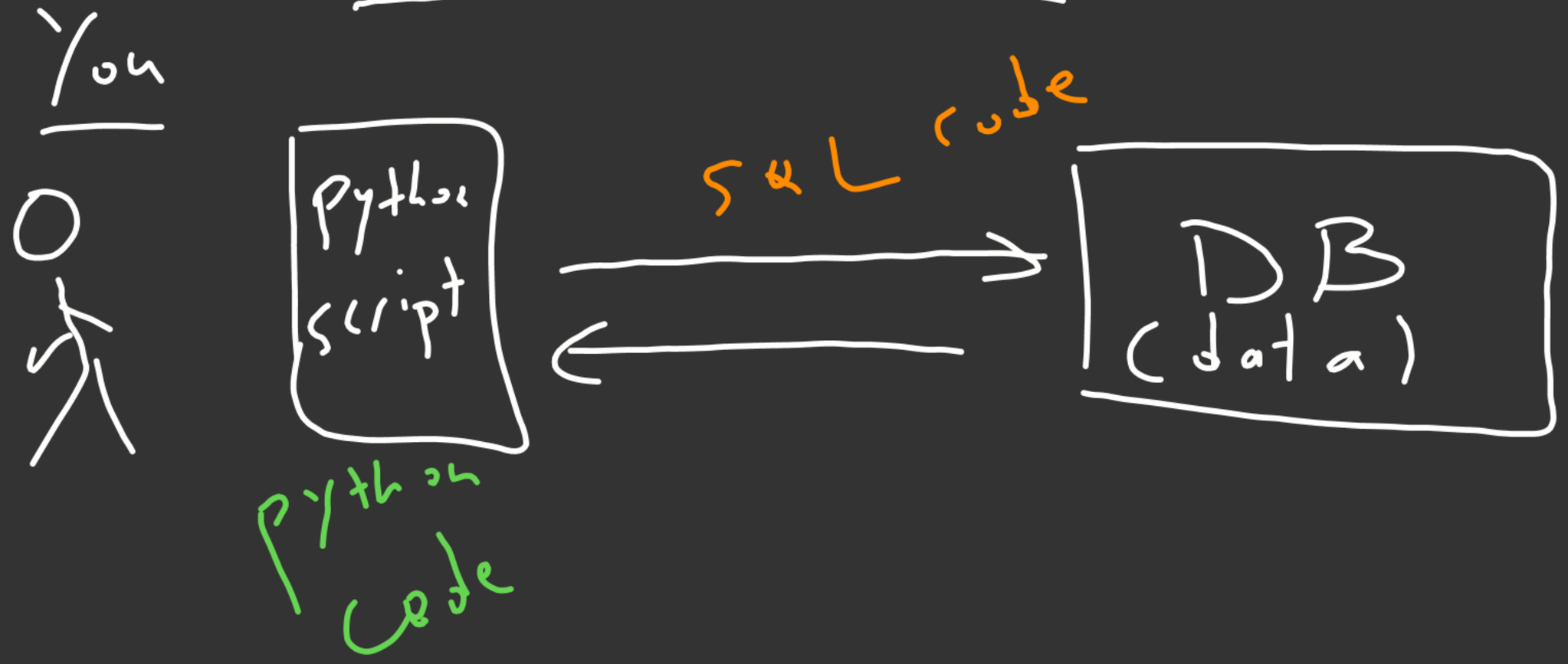
Good morning!

We start at 10:00

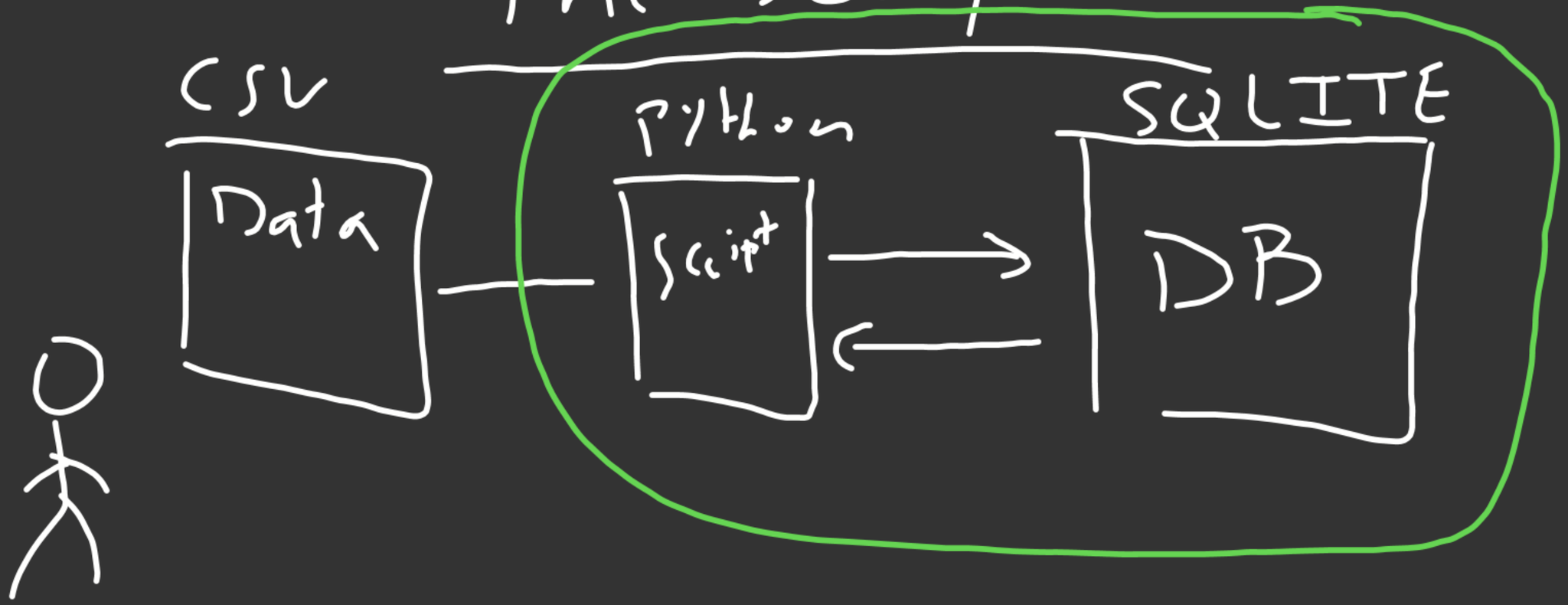
Data Sources

- Excel / CSV / Tabular data sheets
- Databases
- Web

The Setup (1)

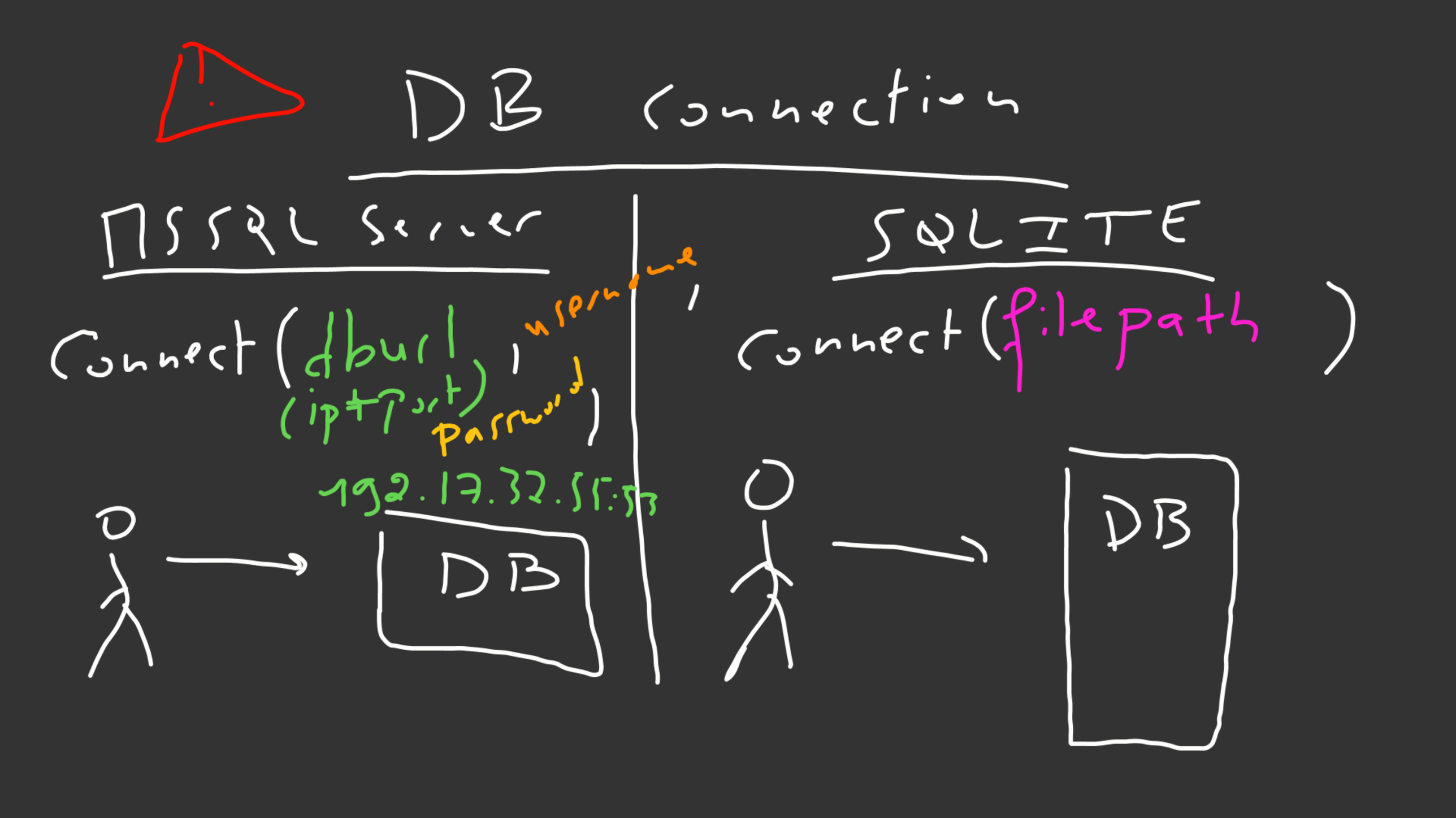


The Setup (2)



SQLITE

- DB is stored on a file
- Lightweight DB
- ✓ No installation or configuration
- ✗ Security issues


DB Connection
MSSQL Server
SQLITE
Connect(dburl, username, password)
(ip+Port)
192.17.32.51:53
DB
Connect(filepath)
DB

Break

Back at 11:15

Feature Engineering

- Fill missing values
- Drop columns
- Apply transformations
- Dimensionality Reduction

Main Steps

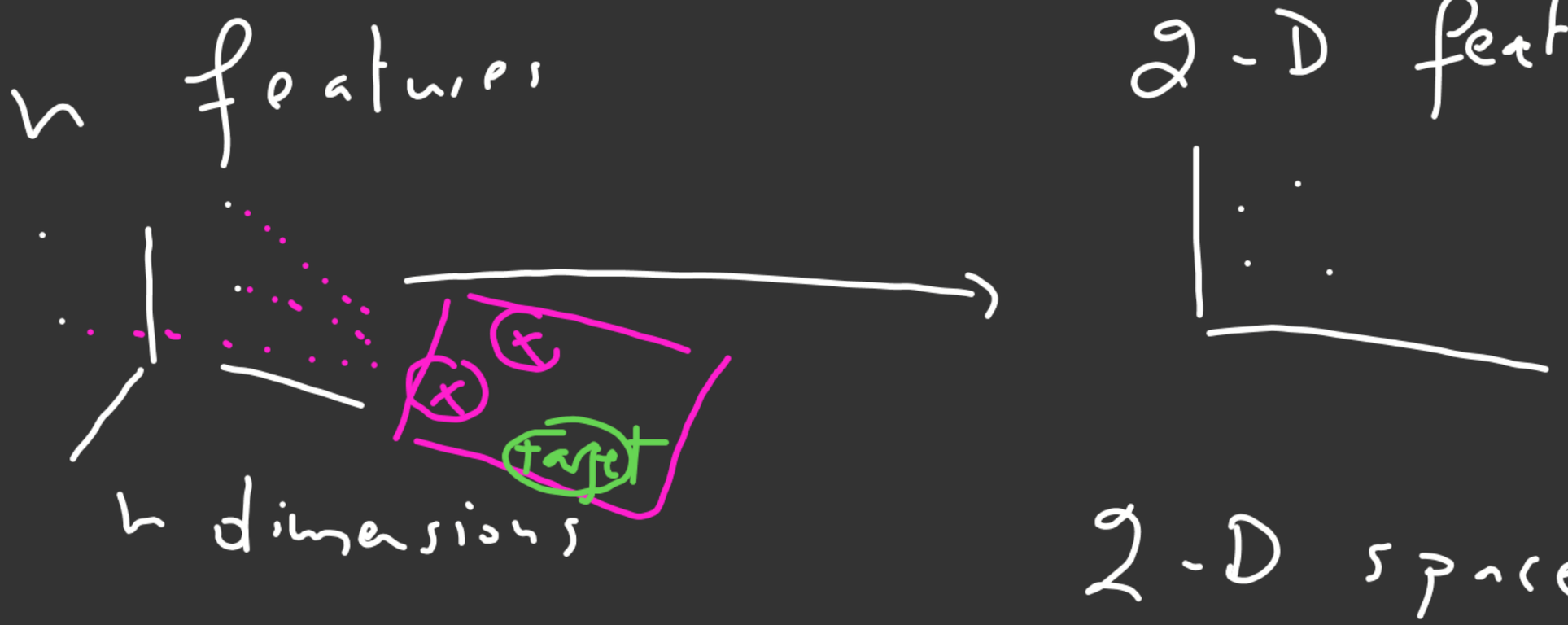
Dimensionality Reduction (1)

Titanic

Cat image

Dimensionality Reduction (2)

Principle Component Analysis (PCA)

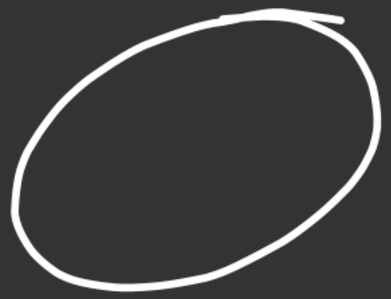


RFE (1)

Feature Selection algorithm
(Recursive Feature Extraction)

RFE (2)

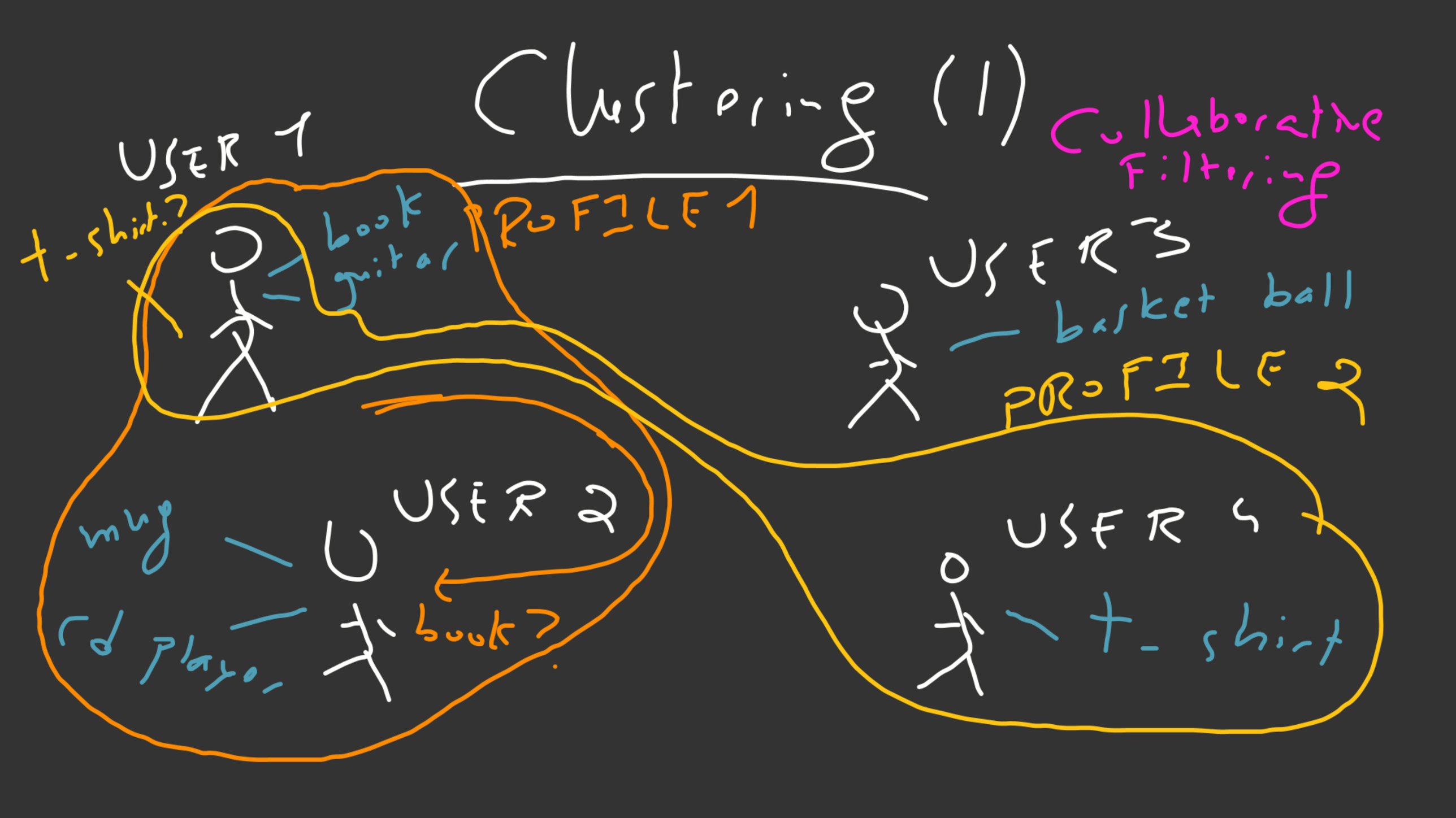
n features model



RFE

+

n-1 features + model


Clustering (1)
Collaborative Filtering
USER 1
t-shirt?
book
guitar
PROFILE 1
USER 3
basket ball
PROFILE 2
USER 2
mug
cd player
book?
USER 4
t-shirt

Clustering (2)

Netflix

USER 1	USER 2	USER 3
- Pirates of the Caribbean	- Rom-com	- Likes Hugh Grant
- Horror movies	- Hates Jim Carrey	

Johnny Depp

Secret Window

Love Actually (Hugh Grant + Rom-Com)

Lunch

Back at 14:00

Filling missing Age (1)

(1, female)

row	Pclass	Sex	Age
0	1	f	38
1	2	f	25
2	1	f	30
3	1	f	

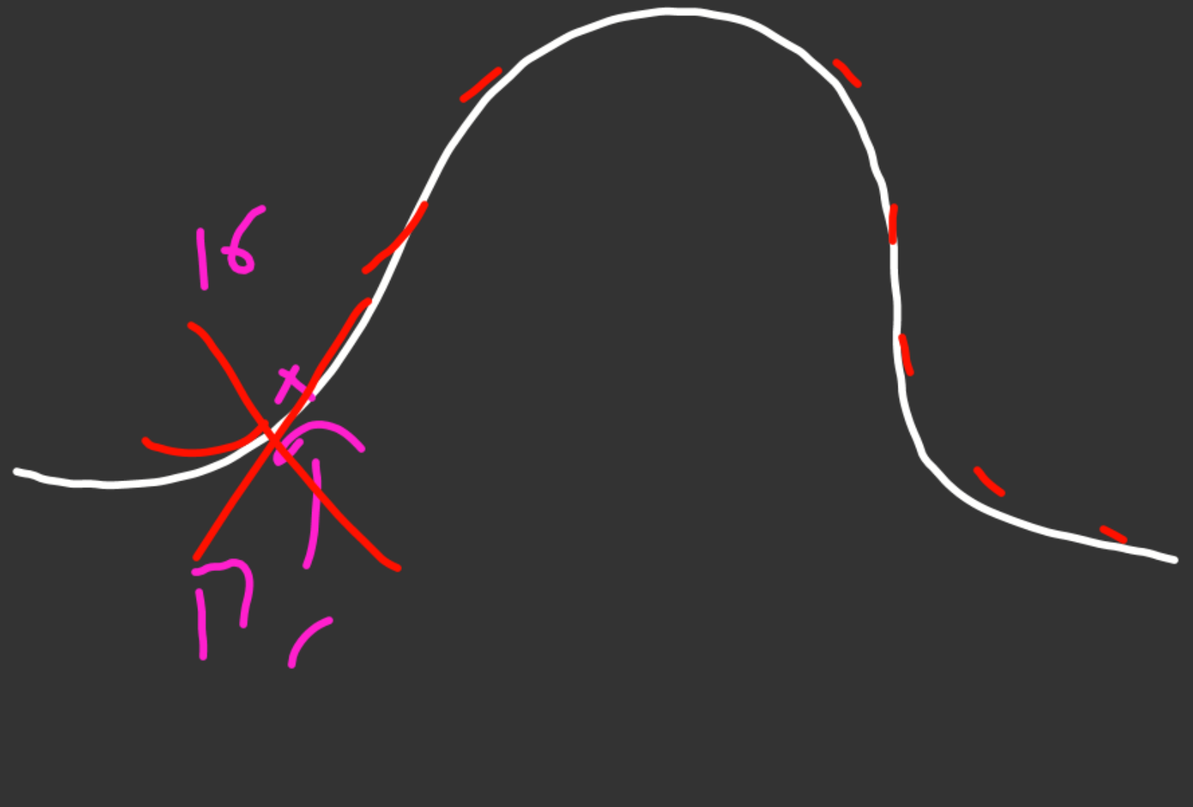
row	Pclass	Sex	Age
0	1	f	38
2	1	f	30
3	1	f	

Filling Missing Ages (2)

row	Relatives	Sex	Age
0	1	f	38
2	1	f	30
3	1	f	X

[3, ...]

Filling Data



Break

Back at 15:05

Features

Numerical

- Discrete
  e.g. 1, 2, 3 ...
- Continuous
  e.g., 1.1, 2.1, 1.001, ...

Categorical

- Nominal
  e.g., Japan, USA, ...
- Ordinal
  e.g., high, medium, low ...
- Free text
  e.g., summary of a book

Encoding (1)

Label

Country

Japan → 1

USA → 2

India → 3

India >> Japan

Pros & Cons

- Numerical ✓
- Good for many values ✓
- Can infer wrong relationships (ordinal) ✗

Encoding (2)

⚠ Leave-one-out = Represent k values with k-1 columns

Vector (one-hot)

Country	J	U	I
Japan	1	0	0
USA	0	1	0
India	0	0	1

Pros & Cons

- No inference of wrong relationships ✓
- Not good for many value representation ✗
- Memory overload ✗

Encoding (3)